Metody obliczeniowe dla wielkoskalowych danych w diagnostyce medycznej

Piotr Dittwald*

autoreferat rozprawy doktorskiej

Analiza bioinformatyczna w genetyce oraz proteomice

Dynamicznie rozwijające się gałęzie nauk o życiu związane są z przetwarzaniem niebagatelnych ilości danych, co wymusza automatyzację znaczącej części analiz. Zaawansowane metody bioinformatyczne stanowią obecnie standardowy krok wielu badań w genetyce i proteomice. Kompleksowa analiza architektury ludzkiego genomu oraz podstawowych funkcjonalnych cząsteczek, jakimi są białka, zmierza do wyjaśnienia zjawisk warunkujących ludzkie życie. W niniejszej rozprawie pokazujemy różnorodne zastosowania metod obliczeniowych w tej niezwykłej naukowej przygodzie.

Metody i wyniki analiz stabilności genomu

W pierwszej części rozprawy zajmujemy się nawracającymi rearanżacjami genomowymi. Są to zmiany strukturalne, wśród których rozważać będziemy delecje, duplikacje oraz inwersje powstałe *de novo* (tj. nieodziedziczone) w tych samych miejscach w genomie u różnych osobników. Obszary genomu, których u danego osobnika jest mniej lub więcej niż w referencyjnym genomie, nazywamy wariantami o zmienionej liczbie kopii (*ang. Copy-Number*

*promotorzy: dr hab. Anna Gambin, dr hab. Paweł Stankiewicz

Variants; CNVs). Głównym mechanizmem odpowiedzialnym za powstawanie nawracających CNVs jest niealleliczna rekombinacja homologiczna (*ang. nonallelic homologous recombination*; NAHR). W mechanizmie tym główną rolę odgrywają długie fragmenty podobnych sekwencji, takie jak sekwencje o niskiej liczbie powtórzeń (*ang. low-copy repeats*; LCRs).

LCRs, inaczej segmentalne duplikacje (SD) [Bailey et al., 2002], są zdefiniowane jako pary sekwencji DNA dłuższe niż 1 kb, o współczynniku podobieństwa sekwencyjnego powyżej 90%. W pracy [Stankiewicz and Lupski, 2002] zostało pokazane, że długie ($10 - 400$ kb) fragmenty LCRs o wysokim (powyżej 97%) współczynniku podobieństwa sekwencyjnego mogą sprzyjać zachodzeniu m.in. inwersji (w przypadku par sekwencji LCRs o orientacji odwrotnej), oraz delecji i duplikacji (sekwencje LCRs o orientacji zgodnej).

Na koniec warto wspomnieć o jednej z głównych technologii biologii molekularnej, wykorzystywanej w niniejszej pracy. Jest nią metoda porównawczej hybrydyzacji genomowej do mikromacierzy (*ang. microarray-based Comparative Genomic Hybridization*; aCGH), pozwalająca na wysokoprzepustową analizę danych [Chial, 2008] i wykrywanie CNVs o rozmiarze dziesiątek tysięcy par zasad (lub dłuższych).

Znane oraz potencjalne nawracające delecje oraz duplikacje

W pracy Dittwald et al. [2013c] przeanalizowaliśmy obszary genomowe związane ze znanymi syndromami skojarzonymi z delecjami lub duplikacjami powodowanymi przez mechanizm NAHR. Ponadto, przeszukaliśmy unikalną kliniczną bazę danych mikromacierzowej analizy chromosomów (*ang. chromosomal microarray analysis*, CMA). Za pomocą podzbioru sekwencji LCRs szczególnie sprzyjających powstawaniu nawracających delecji i duplikacji (oznaczonych jako DP-LCRs; *ang. directly oriented paralogous LCRs*), skonstruowaliśmy (za pomocą metody hierarchicznej) klastry sekwencji LCRs, oraz przedstawiliśmy całogenomową mapę regionów podatnych na niestabilność uwarunkowaną przez NAHR. W trakcie naszych badań opisaliśmy również w regionie 2q12.2q13 nowe zespoły nawracających delecji, korelując z nimi dane pacjentów.

Wiele rearanżacji z bazy CMA udało nam się skojarzyć ze znanymi zespołami, co umożliwiło ustalenie skłonności do występowania poszczególnych jednostek chorobowych w analizowanej populacji pacjentów. Kolejny krok

analizy dotyczył badania częstości rearanżacji powstających *de novo*. Do opisania związku między cechami architektury genomu (dotyczącymi zarówno DP-LCRs jak i klastrów LCRs) a częstością delecji *de novo*, wykorzystane zostało modelowanie statystyczne (korelacja rangowa Spearmana oraz model regresji Poissona). W wyniku przeprowadzonych analiz okazało się, że częstość badanych rearanżacji koreluje negatywnie z odległością par DP-LCRs, oraz koreluje pozytywnie ze współczynnikiem podobieństwa sekwencji.

Dodatkowo, wyodrębniliśmy geny narażone na uszkodzenie przez mechanizm NAHR, ze szczególnym uwzględnieniem genów wrażliwych na dawkę (*ang. dosage sensitive genes*) oraz genów skojarzonych z chorobami w bazie danych OMIM (`http://www.omim.org/`).

Schemat powyższych analiz przedstawiony jest na Rycinie 1.

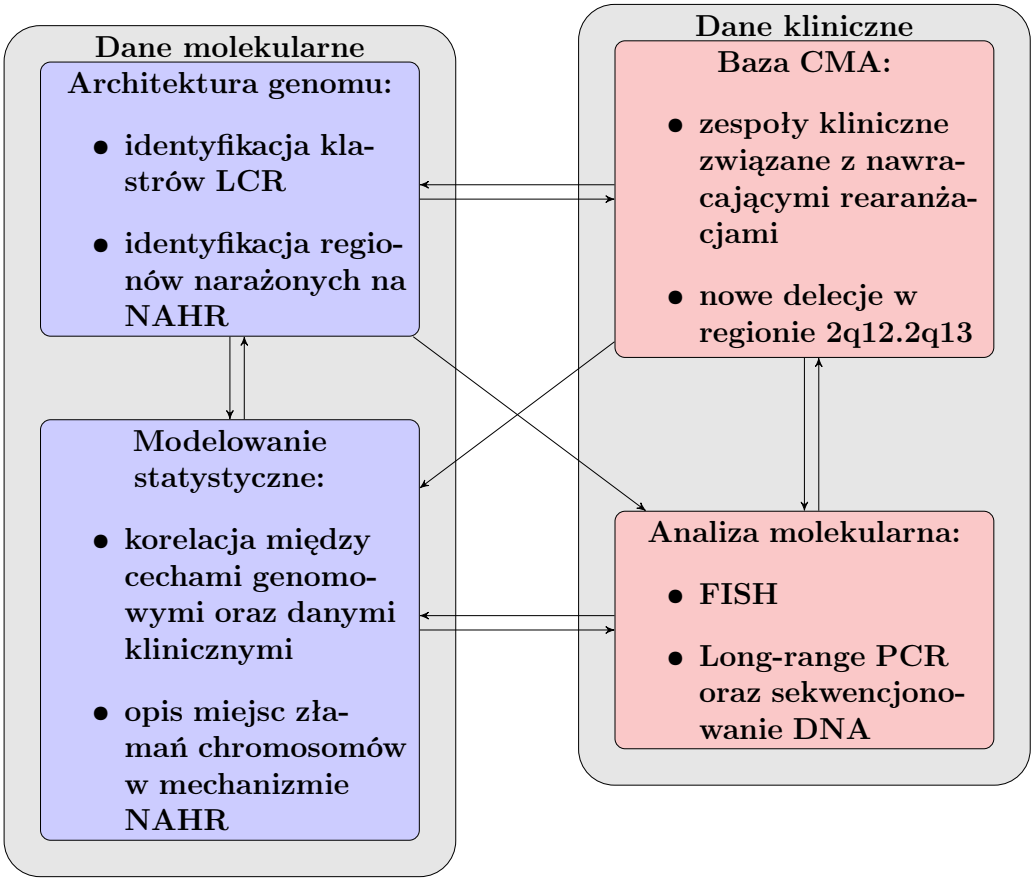
Nawracające inwersje

Technologia aCGH nie umożliwia wykrywania zrównoważonych rearanżacji genomowych, jakimi są odwrócone fragmenty sekwencji DNA, tj. inwersje. Jest to prawdopodobnie przyczyną opisania do tej pory stosunkowo niewielkiej liczby patogennych przypadków nawracających inwersji.

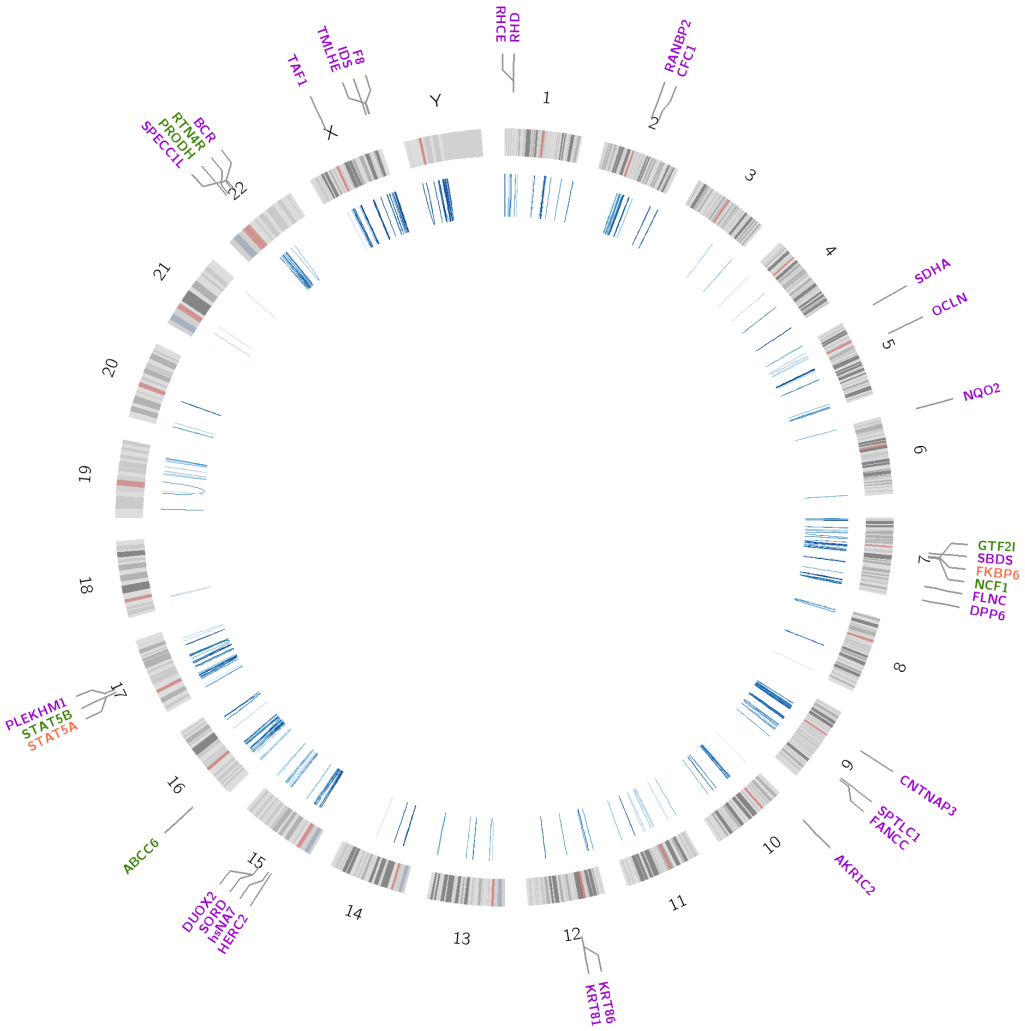
W pracy Dittwald et al. [2013b] wyodrębniliśmy podzbiór sekwencji IP-LCRs (*inversely oriented paralogous LCRs*), który jest szczególnie podatny na pośredniczenie w powstawaniu nawracających inwersji oraz zaproponowaliśmy mapę potencjalnie niestabilnych regionów genomu. Ponadto, przeanalizowaliśmy geny, które mogą być uszkodzone w wyniku tychże rearanżacji, koncentrując się na genach wrażliwych na dawkę oraz skojarzonych ze znanymi jednostkami chorobowymi (por. Ryc. 2). Dodatkowo, przeanalizowaliśmy inwersje z bazy danych DGV (*Database of Genomic Variants*)[Zhang et al., 2006] zawierającej CNVs pochodzące od zdrowych osób. W wyniku naszych analiz zidentyfikowaliśmy inwersje, które mogły być spowodowane przez mechanizm NAHR, szacując także ich statystyczną istotność.

Metody analizy proteomicznej

Atomy chemiczne zbudowane są z pozytywnie naładowanych protonów, nieposiadających ładunku neutronów, oraz negatywnie naładowanych elektronów. Protony i neutrony, zwane także nukleonami, tworzą jądro atomowe, gdzie skupia się przeważająca część masy całego atomu (w związku z tym



Rycina 1: Analiza z pracy Dittwald et al. [2013c]. Kolory fioletowy i różowy oznaczają, odpowiednio, przetwarzanie danych molekularnych oraz klinicznych. Strzałki wskazują przepływ danych dokonany lub wspierany za pomocą zaproponowanych zautomatyzowanych procedur. Na podstawie oryginalnego rysunku dostarczonego przez dr. Annę Gambin.



Rycina 2: Rysunek wykonany przy użyciu programu Circos [Krzywinski et al., 2009], ilustrująca podzbiór zidentyfikowanych genów, które są podatne na uszkodzenie przez inwersje powodowane przez mechanizm NAHR. Kolorami oznaczono geny skojarzone z jednostkami chorobowymi (fiolet), wrażliwe na dawkę (czerwień), oraz należące do obydwu tych grup (zieleń). Źródło: Dittwald et al. [2013b].

w naszych analizach nie rozważmy wpływu elektronów na masę cząsteczek). Pierwiastki chemiczne mają swoje stabilne warianty izotopowe, różniące się liczbą neutronów. Na potrzeby niniejszych rozważań ograniczymy się do pięciu pierwiastków chemicznych budujących białka: węgla (C), wodoru (H), azotu (N), tlenu (O) oraz siarki (S). Najlżejsze warianty izotopowe pierwiastków nazywamy wariantami monoizotopowymi (w naszym przypadku są to $^{12}C, ^{1}H, ^{14}N, ^{16}O, ^{32}S$). Podstawową jednostką masy używaną przez nas będzie dalton (Da), zdefiniowany jako $\frac{1}{12}$ masy węgla ^{12}C, co odpowiada w przybliżeniu 1.66×10^{-27} kg. Analizowane pierwiastki chemiczne mają, odpowiednio, dwa (węgiel: $^{12}C, ^{13}C$; wodór: $^{1}H, ^{2}H$; azot: $^{14}N, ^{15}N$), trzy (tlen: $^{16}O, ^{17}O, ^{18}O$), oraz cztery (siarka: $^{32}S, ^{33}S, ^{34}S, ^{36}S$) stabilne izotopy. Każdy z nich ma swoją masę, oznaczoną jako $M_{C_{12}}, \ldots, M_{S_{36}}$, oraz występuje w środowisku z pewnym prawdopodobieństwem (odpowiednio, $P_{C_{12}}, \ldots, P_{S_{36}}$).

Spektrometria masowa to jedna z najpopularniejszych obecnie metod analitycznych wykorzystywanych w proteomice do badania składu mieszanin, wnosząca niebagatelny wkład w rozwój badań nad działaniem układów biologicznych [Cravatt et al., 2007, Chandramouli and Qian, 2009]. Aparatura wykorzystywana do analizy spektrometrycznej, czyli spektrometr masowy, składa się z trzech głównych komponentów: (1.) źródła jonizacji – cząsteczki są tu ładowane, tworząc jony, a następnie umieszczane w fazie gazowej; (2.) analizatora masy – jony są rozdzielane pod względem wartości współczynnika stosunku masy do ładunku (m/z); (3.) detektora – tworzone jest widmo sygnałów, które wartościom współczynnika (m/z) przypisuje częstość występowania.

Rozkłady izotopowe cząsteczek

Rozważmy cząsteczkę $\xi(v, w, x, y, z)$ złożoną z v atomów węgla, w atomów wodoru, x atomów azotu, y atomów tlenu oraz z atomów siarki. Dla uproszczenia będziemy także używać oznaczenia ξ, gdy parametry $v, \ldots, z$ w sposób oczywisty wynikają z kontekstu.

Analogicznie jak w przypadku atomów, także cząsteczki mają swoje warianty izotopowe. Każdy wariant ma swoją masę oraz odpowiadające jej prawdopodobieństwo, będące odpowiednio, sumą mas oraz iloczynem prawdopodobieństw składających się nań wariantów izotopowych atomów.

Najlżejszy wariant izotopowy cząsteczki (składający się wyłącznie z monoizotopowych wariantów atomów) nazywamy jej wariantem monoizotopo-

wym. Jego masa (zwana masą monoizotopową), dla cząsteczki ξ, wynosi:

$$M_{mono} = vM_{C_{12}} + wM_{H_1} + xM_{N_{14}} + yM_{O_{16}} + zM_{S_{32}},$$

a odpowiadające jej prawdopodobieństwo:

$$P_{mono} = P^v_{C_{12}} \times P^w_{H_1} \times P^x_{N_{14}} \times P^y_{O_{16}} \times P^z_{S_{32}}.$$

Naszą analizę możemy przeprowadzić na różnym poziomie przybliżenia. W podejściu dokładnym rozróżniamy każde dwa warianty różniące się masą. Jednakże, w przypadku nawet małych cząsteczek, liczba dokładnych wariantów izotopowych bardzo szybko rośnie i staje się trudna do dokładnego przeanalizowania. Dlatego w praktyce dobrym podejściem staje się rozważanie wariantów zagregowanych pod względem liczby dodatkowych neutronów względem wariantu monoizotopowego. Średnią masę wariantu zagregowanego obliczamy jako średnia ważoną składających się nań wariantów dokładnych.

Wyniki analizy proteomicznej

Zagregowane warianty izotopowe

Celem w tej części analizy jest efektywne przetwarzanie zagregowanych rozkładów izotopowych. Oznaczmy jako q_j prawdopodobieństwo j-tego zagregowanego wariantu dla cząsteczki ξ, które może być obliczone ze wzoru:

$$q_j = \sum_k p_{jk} \tag{1}$$

oraz średnią masę tegoż wariantu zdefiniowaną jako:

$$E(m_j) = \bar{m_j} = \frac{\sum_k m_{jk} p_{jk}}{\sum_k p_{jk}}. \tag{2}$$

Przez m_{jk} oraz p_{jk} oznaczamy, odpowiednio, masy i prawdopodobieństwa wariantów dokładnych (indeksowanych za pomocą k) przynależących do j-tego wariantu zagregowanego.

Opracowany przez nas algorytm o nazwie BRAIN (**B**affling **R**ecursive **A**lgorithm for **I**sotopic distributio**N** calculations), przedstawiony w pracy Claesen et al. [2012], oblicza zagregowany rozkład izotopowy dla cząsteczki

$C_vH_wN_xO_yS_z$, wykorzystując dwie funkcje tworzące wyrażone w postaci wielomianów. Pierwszy z nich, Q, jest zdefiniowany jako:

$$\begin{aligned} Q(I;v,w,x,y,z) = &\left(P_{C_{12}}I^0 + P_{C_{13}}I^1\right)^v \quad \times \\ &\left(P_{H_1}I^0 + P_{H_2}I^1\right)^w \quad \times \\ &\left(P_{N_{14}}I^0 + P_{N_{15}}I^1\right)^x \quad \times \\ &\left(P_{O_{16}}I^0 + P_{O_{17}}I^1 + P_{O_{18}}I^2\right)^y \quad \times \\ &\left(P_{S_{32}}I^0 + P_{S_{33}}I^1 + P_{S_{34}}I^2 + P_{S_{36}}I^4\right)^z \quad . \end{aligned}$$

Druga funkcja, U, wyraża się za pomocą wielomianu Q w następujący sposób:

$$\begin{aligned} &U(I;v,w,x,y,z) = \\ &vQ(I;v-1,w,x,y,z)\left(P_{C_{12}}M_{C_{12}} + P_{C_{13}}M_{C_{13}}I^1\right) \\ &+wQ(I;v,w-1,x,y,z)\left(P_{H_1}M_{H_1} + P_{H_2}M_{H_2}I^1\right) \\ &+xQ(I;v,w,x-1,y,z)\left(P_{N_{14}}M_{N_{14}} + P_{N_{15}}M_{N_{15}}I^1\right) \\ &+yQ(I;v,w,x,y-1,z)\left(P_{O_{16}}M_{O_{16}} + P_{O_{17}}M_{O_{17}}I^1 + P_{O_{18}}M_{O_{18}}I^2\right) \\ &+zQ(I;v,w,x,y,z-1) \times \\ &\left(P_{S_{32}}M_{S_{32}} + P_{S_{33}}M_{S_{33}}I^1 + P_{S_{34}}M_{S_{34}}I^2 + P_{S_{36}}M_{S_{36}}I^4\right) \ . \end{aligned}$$

Algorytm oblicza kolejno współczynniki w obu funkcjach, używając algebraicznej teorii opartej o tożsamości Newtona-Girarda oraz wzory Viète'a [Séroul, 2000, Vinberg, 2003]. W szczególności, otrzymujemy następującą iteracyjną formułę na prawdopodobieństwa wariantów zagregowanych:

$$q_j = -\frac{1}{j}\sum_{l=1}^{j} q_{j-l}\psi_l,$$

gdzie ψ_l jest sumą $(-l)$-tych potęg pierwiastków wielomianu $Q(I;v,w,x,y,z)$.

Ponadto, zaimplementowaliśmy (w języku R) algorytm BRAIN w pakiecie o tej samej nazwie [Dittwald et al., 2013a] jako część repozytorium o nazwie Bioconductor [Gentleman et al., 2004]. Zademonstrowaliśmy także użyteczność naszej metody przy wyskoprzepustowym przetwarzaniu proteomicznej bazy danych (na przykładzie bazy Uniprot), konstruując model liniowy wyznaczający masę monoizotopową na podstawie średniej masy najczęstszego zagregowanego wariantu izotopowego. Tego typu podejście może być potencjalnie wykorzystane przez eksperymentatorów, którzy nie są w

stanie zaobserwować masy monoizotopowej dla dużych peptydów, a chcieliby ją użyć w celu identyfikacji cząsteczki. Dodatkowo, w ramach naszych badań przeprowadziliśmy analizę implementacji algorytmu BRAIN w języku C++ [Hu et al., 2013].

W dalszej kolejności, zaproponowaliśmy algorytm BRAIN 2.0 [Dittwald and Valkenborg, 2014]. Obejmuje on dwa ulepszenia pozwalające na poprawienie złożoności czasowej i pamięciowej przy liczeniu proporcji pomiędzy prawdopodobieństwami sąsiednich zagregowanych wariantów izotopowych, oraz alternatywną metodę obliczania sum potęgowych pierwiastków wielomianów, pozwalającą uniknąć liczenia wartości poszczególnych pierwiastków.

W celu ramach przykładu praktycznej analizy danych wielkoskalowych, proponujemy zautomatyzowaną procedurę do rozróżniania pomiędzy sygnałami pochodzącymi od peptydów oraz od lipidów. Skonstruowaliśmy zestaw klasyfikatorów metodą lasów losowych, biorąc pod uwagę różne parametry przy modelowaniu rozdzielczości oraz szumu aparatury. Przeprowadzając eksperymenty, wykorzystaliśmy bazy danych z formułami chemicznymi, jak również rzeczywistą mieszaninę lipidów i peptydów przeanalizowaną za pomocą spektrometru masowego.

Dokładne rozkłady izotopowe

Następny etap naszej analizy obejmuje próbę opisu dokładnej struktury izotopowej dla konkretnych (np. najczęstszych) zagregowanych wariantów izotopowych. W tym celu wprowadzamy funkcję:

$$Q^{\perp}(I, J, K; v, w, x, y, z) = \sum_j (\sum_k p_{jk} J^{m_{jk}} K^{m_{jk}}) I^j,$$

a następnie udowadniamy, że wariancję zagregowanych wariantów izotopowych można otrzymać obliczając współczynniki następującego wielomianu:

$$\frac{\partial^2}{\partial J \partial K} Q^{\perp}(I, J, K; v, w, x, y, z)|_{J=K=1}.$$

Ponadto, (teorio-informacyjna) entropia dla j-tego zagregowanego wariantu (oznaczona jako $H(j)$) może być obliczona przy użyciu wielomianowej funkcji tworzącej oraz następującego równania:

$$H(j) = \frac{-\sum_k p_{jk} \log(p_{jk})}{\sum_k p_{jk}} + \log(\sum_k p_{jk}).$$

Po przetworzeniu bazy danych Uniprot zbudowaliśmy model liniowy przewidujący wariancję najczęstszego zagregowanego wariantu na podstawie jego średniej masy. Następnie oszacowaliśmy rozrzut pomiędzy najcięższym a najlżejszym dokładnym wariantem j-tego zagregowanego wariantu izotopowego, ograniczając go z góry przez:

$$j \cdot (\mu_{2H} - \mu_{15N}),$$

co było przydatne w przewidywaniu nachodzenia na siebie sąsiednich wariantów izotopowych.

Artykuły i manuskrypty w przygotowaniu

Pierwsza część rozprawy dotycząca genetyki oparta jest na następujących artykułach:

- Piotr Dittwald*, Tomasz Gambin* et al (2013). NAHR-mediated copy-number variants in a clinical population: Mechanistic insights into both genomic disorders and Mendelizing traits. (* wkład równomierny) Genome Research 23, 9: 1395-409,

- Piotr Dittwald*, Tomasz Gambin*, Claudia Gonzaga-Jauregui*, Claudia M.B. Carvalho, James R. Lupski, Paweł Stankiewicz, Anna Gambin (2013). Inverted low-copy repeats and genome instability – a genome-wide approach, Human Mutation, 34, 1: 210-20. (* wkład równomierny).

Zawartość drugiej, proteomicznej, części rozprawy oparta jest na artykułach opublikowanych:

- Jürgen Claesen*, Piotr Dittwald*, Dirk Valkenborg, Tomasz Burzykowski (2012). An efficient method to calculate the aggregated isotopic distribution and exact center-masses, Journal of the American Society for Mass Spectrometry;23(4): 753-63. (* wkład równomierny)

- Piotr Dittwald, Jürgen Claesen, Tomasz Burzykowski, Dirk Valkenborg, Anna Gambin (2013). BRAIN: a universal tool for high-throughput calculations of the isotopic distribution for mass spectrometry. Analytical Chemistry, 85, 4: 1991-4

- Piotr Dittwald, Dirk Valkenborg (2014). BRAIN 2.0: Time and Memory Complexity Improvements in the Algorithm for Calculating the Isotope Distribution. Journal of the American Society for Mass Spectrometry, 25(4): 588-94,

- Han Hu*, Piotr Dittwald*, Joseph Zaia, Dirk Valkenborg (2013). Comment on Ćomputation of isotopic peak center-mass distribution by Fourier Transform"(* wkład równomierny). Analytical Chemistry, 85(24): 12189-92.

oraz manuskrytpach będących w przygotowaniu:

- Piotr Dittwald, Vu Trung Nghia, Glenn A. Harris, Richard M. Caprioli, Raf Van de Plas, Kris Laukens, Anna Gambin, Dirk Valkenborg, Towards automated discrimination of lipids versus peptides from full scan mass spectra,

- Piotr Dittwald, Jürgen Claesen, Dirk Valkenborg, Alan L. Rockwood, Anna Gambin, On isotopic fine structure distribution and limits to resolution in mass spectrometry.

Literatura

J. A. Bailey, Z. Gu, R. A. Clark, K. Reinert, R. V. Samonte, S. Schwartz, M. D. Adams, E. W. Myers, P. W. Li, and E. E. Eichler. Recent segmental duplications in the human genome. *Science*, 297(5583):1003–1007, 2002.

K. Chandramouli and P. Y. Qian. Proteomics: challenges, techniques and possibilities to overcome biological sample complexity. *Human Genomics and Proteomics*, 2009, 2009.

H. Chial. Cytogenetic methods and disease: Flow cytometry, CGH, and FISH. *Nature Education*, 1(1), 2008.

J. Claesen, P. Dittwald, T. Burzykowski, and D. Valkenborg. An efficient method to calculate the aggregated isotopic distribution and exact center-masses. *Journal of the American Society for Mass Spectrometry*, 23:753–763, 2012.

B. F. Cravatt, G. M. Simon, and J. R. Yates. The biological impact of mass-spectrometry-based proteomics. *Nature*, 450(7172):991–1000, 2007.

P. Dittwald and D. Valkenborg. BRAIN 2.0: Time and Memory Complexity Improvements in the Algorithm for Calculating the Isotope Distribution. *Journal of the American Society for Mass Spectrometry*, 25(4):588–594, 2014.

P. Dittwald, J. Claesen, T. Burzykowski, D. Valkenborg, and A. Gambin. BRAIN: a universal tool for high-throughput calculations of the isotopic distribution for mass spectrometry. *Analytical Chemistry*, 85:1991–1994, 2013a.

P. Dittwald, T. Gambin, C. Gonzaga-Jauregui, C. M. Carvalho, J. R. Lupski, P. Stankiewicz, and A. Gambin. Inverted low-copy repeats and genome instability–a genome-wide analysis. *Human Mutation*, 34(1):210–220, 2013b.

P. Dittwald, T. Gambin, P. Szafranski, J. Li, S. Amato, M. Y. Divon, L. X. Rodriguez Rojas, L. E. Elton, D. A. Scott, C. P. Schaaf, W. Torres-Martinez, A. K. Stevens, J. A. Rosenfeld, S. Agadi, D. Francis, S. H. Kang, A. Breman, S. R. Lalani, C. A. Bacino, W. Bi, A. Milosavljevic, A. L. Beaudet, A. Patel, C. A. Shaw, J. R. Lupski, A. Gambin, S. W. Cheung, and P. Stankiewicz. NAHR-mediated copy-number variants in a clinical population: mechanistic insights into both genomic disorders and Mendelizing traits. *Genome Research*, 23(9):1395–1409, 2013c.

R. C. Gentleman, J. V. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. Yang, and J. Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 5(10):R80, 2004.

H. Hu, P. Dittwald, J. Zaia, and D. Valkenborg. Comment on the computation of isotopic peak center-mass distribution by fourier transform. *Analytical Chemistry*, 85:12189–12192, 2013.

M. Krzywinski, J. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S. J. Jones, and M. A. Marra. Circos: an information aesthetic for comparative genomics. *Genome Research*, 19(9):1639–1645, 2009.

R. Séroul. *Programming for Mathematicians*. Berlin: Springer-Verlag, 2000.

P. Stankiewicz and J. R. Lupski. Genome architecture, rearrangements and genomic disorders. *Trends in Genetics*, 18(2):74–82, 2002.

E. B. Vinberg. *A course in algebra*. American Mathematical Society, Providence, 2003.

J. Zhang, L. Feuk, G. E. Duggan, R. Khaja, and S. W. Scherer. Development of bioinformatics resources for display and analysis of copy number and other structural variants in the human genome. *Cytogenetic and Genome Research*, 115(3-4):205–214, 2006.